DOCUMENT RESUME

ED 224 022 CS 207 264

AUTHOR Houston, Robert
TITLE Faculty Evaluation of Standardized Tests of Writing Ability.
PUB DATE [81]
NOTE 17p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College English; English Departments; Essay Tests; Higher Education; Item Analysis; Objective Tests; *Standardized Tests; *Student Placement; Test Construction; *Testing Problems; *Test Reliability; Test Selection; Test Use; *Test Validity; *Writing Evaluation; Writing Instruction; Writing Skills

ABSTRACT

Standardized tests of writing ability have individual and shared limitations and deficiencies that should be acknowledged by test designers and users. Most institutions use the portions of standardized tests that test ability to proofread and edit, but they do not use the optional essay sections that actually require students to write. To assure validity of a particular test requires item analysis by the department considering using it. An objective test of the student's mastery of standard, edited English does not test equally important abilities to choose a topic, evolve a thesis statement, and actually write a unified, coherent essay. Some teachers will not accept objective tests, insisting instead on writing samples. Other educators claim that essay tests lack reliability and do not correlate with objective test scores and course grades. Work by the Educational Testing Service and College Entrance Examination Board researchers shows how these problems can be overcome. College English departments should conduct score gains studies to give credibility to claims of content validity. Since testing services often do not or cannot give enough information on item analysis, score gains, and correlation in informational booklets, and since individual departments differ from each other, every English department must correlate the composition grades and test scores of its students. (JL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED224022

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert Houston

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Faculty Evaluation of Standardized Tests of Writing Ability

Robert Houston

Of the 1,294 San Francisco State College freshmen who enrolled the fall of 1971, nearly half did not have to take freshman English. Earning a minimum score of 429 (25th percentile) on the original College Level Examination Program (CLEP) General Examination in English Composition exempted 531 of them from freshman English.[1] These students should have been exempted only if the test were an accurate measure of writing ability and if the cutoff score of 429 identified students not needing the course work. Unfortunately, this test and all other standardized tests of writing ability have their individual and shared limitations and deficiencies. What these limitations and deficiencies are should be acknowledged by the test writers in their informational booklets on the test. Or, as necessary, they should be detected by the test users themselves. Only the ignorant, ill-informed, or irresponsible will use a test without investigating its validity, its norming, its scoring, and the like.

Before San Francisco State and nearly three hundred other colleges used the original Examination to exempt students,[2] each institution should have investigated the content of the test and should have learned the meaning of the test scores. Educational Testing Service (ETS), the producer of the test, and the College Entrance Examination Board (CEEB), did provide information about the Examination in two booklets; however, the ETS-CEEB consortium either superficially acknowledged or ignored deficiencies in the content and scoring of the

CS207264

test. The status and power of ETS-CEEB could have intentionally or unintentionally convinced a naive and trusting test user that a student who scored 429 on the Examination would not benefit from taking freshman English. Sophisticated and skeptical test users like Caldwell, Rudolph and Summers, and Archer and Nickens were not convinced.

In his investigation of the content of the Examination, Edward Caldwell had an item analysis conducted by"subject matter specialists." As a result of their analysis, they judged nearly one-fourth of the items to be high school level.[3] Since Caldwell provides neither the credentials of his specialists nor the criteria for distinguishing high school from college level course work, his judgment is suspect. But the point that he makes about the necessity of investigating the content of a test through item analysis is not. A test of a college level course must have items representative of that level of course work if it is to have content validity.

Robert S. Rudolph and Richard M. Summers also conducted an item analysis of the Examination. Unlike Caldwell, they did not question its content based on their analysis. Instead, they ask if a test can have content validity for a composition course when the student is not asked to write.[4]

According to ETS-CEEB's informational booklets, the scale scores and percentile ranks for the Examination were based on the performance of a sample of 2,582 sophomores who took the test in 1963.[5] The scale score of 429/25th percentile used by San Francisco State as its cutting score for exempting students from freshman English was not, as ETS-CEEB concedes, a score indicating mastery. Like any scale score, it was a score indicating the student's position in relation to the other students who took the test (CLEP Scores, p. 7).

The standard error of measurement for the original Examination was 31 (CLEP Scores, p. 13). A student whose scale score was 429 had a "true" score of 398-460 and a "true" percentile rank from the 17th to the 36th. Caldwell converted the scale score to a raw score. He learned that to score 429, a student would only have had to have answered approximately one-third of the 100 items correctly and none incorrectly (p. 700).

If the scale score was not an indicator of mastery and if the standard error of measurement and the raw score cast further doubt on using 429 as the cutting score for exemption, why did San Francisco State use it? Like other institutions, San Francisco State accepted the ETS-CEEB recommendation, one endorsed by the American Council of Education (ACE), to use the 25th percentile (CLEP Scores, p. 46). That recommendation should have been supported by correlation studies of course grades and test scores which ETS-CEEB and ACE should have conducted to show that 429 would be the score of a student who at least passed freshman English. Amazingly, ETS-CEEB and ACE reported no such studies.

If San Francisco State had conducted an institutional correlation study, its discovery might have been comparable to J. Andrew Archer and Harry C. Nickens': students scoring at the 25th percentile were typically C and D students; students at the 50th percentile were typically A and B students.[6] Although not irrecusable, correlations are at least helpful when ascertaining content validity and cutting scores. Not knowing the correlation that exists between whatever the test is a measure of and whatever the course grade is a measure of precludes an attempt to predict student performance.

In each institution that used the original Examination the faculty

in the department most able to evaluate the students' writing should have been the ones to select the means and criteria for judging student composition ability. The college's English department faculty should appraise the students' composition ability when they are certified competent writers, exempted from composition courses, or placed in remedial, regular, or honors sections of freshman English. Major curriculum decisions affecting student literacy, such as certification, exemption, and placement, should be primarily the English department's, not an administrator's or a college curriculum committee's.

And if a standardized test of writing ability is one of the means used to evaluate student composition ability, the department must examine its validity, norming, scoring, and so on. The testing service's description of the test as well as any institution's research done on the test must be competently analyzed, interpreted, and assessed. The faculty cannot be oblivious to the research, cannot presume candor on the part of the testing service, and cannot deny their duty in determining test acceptability. To do otherwise would be to invite a travesty of test use like that which occurred at San Francisco State and other colleges.

II

Content Validity

Standardized tests like the original CLEP General Examination in English Composition continue to be used by colleges and universities for certifying competence, placement, and exemption. These tests include the American College Testing Program (ACT) English Usage Test, the ETS Test of Standard Written English, the revised CLEP General Examination in English Composition, the Houghton Mifflin College English Placement Test, as well as several others.

Typically these are multiple-choice, 30-45 minute tests having 50-100 items. From 20 to 40 percent of the test items ask the student to make decisions about topic selection and essay and paragraph unity and organization. Usually about two-thirds of the items ask the student to recognize sentence faults, errors in grammar, spelling, punctuation, and capitalization, mistakes in diction, and flaws in style. In short, testing the student's ability to proofread and edit based on his or her knowledge of the proscriptions and conventions of standard edited English is its major purpose. Only when the optional essay section of such a test is required will the student write. Optimally all institutions would use both sections; actually only a very few are willing to confront the difficulties of administering and scoring essay examinations.

A test must measure what it purports to measure if it is to have validity. To make a prediction about student performance in a college course based on the student's test performance, the test items must be representative of the content of that course. To assure a user that the test has content validity, the testing service in its informational booklets for administrators, faculty, and students should provide an analysis of each item's type, the number of each type, and the difficulty of each item. Also, the service should identify the skills needed to answer the test items correctly. Instead of providing the necessary analysis, the testing service may base, but not limit, its claim for content validity on its contention that the writers of its tests include the service's own test specialists and university faculty. Presumably these writers know the content of college composition courses and as a result write test items representative of the content of those courses.

The test writers' credentials may be impressive, but without an item analysis the department cannot presume the test has content validity. Differences in content among freshman composition courses necessitate each department's analyzing every test under consideration for use and comparing its content to the content of the department's courses. The typical test is probably inappropriate for student placement in and exemption from the freshman English course which is primarily a study of literature and in which composition ability plays a secondary role in determining course grades. It may also be inappropriate for students at a private liberal arts college having a highly selective admissions policy. These students may be competent writers whose course work would concentrate on rhetoric and style, not usage errors and sentence faults. If, after securing copies of the test and analyzing each item, the department faculty can say the content of the test, its difficulty, and its emphasis parallel the content of the department's composition courses, then the test has content validity.

Tests, like composition courses, do differ in their emphases and in their demands on the student's knowledge and skills, but nearly all tests focus on the student's mastery of standard edited English. Although included in such tests, items testing the student's ability to recognize proficiently written essays and paragraphs occur far less frequently than items testing the student's ability to proofread and edit sentences. Most standardized tests of writing ability do have content validity inasmuch as the knowledge and skills they test for are among the concerns of most freshman composition teachers, and which, consequently, are part of their courses. Similarities among syllabi, textbooks, and the tests themselves attest to their having content validity for that part of the course work.

Mastering standard edited English and gaining the ability to recognize competently written sentences, paragraphs, and essays are important goals most composition teachers have for their students. Equally important to them is the student's ability to choose a topic, evolve a thesis statement, and actually write a unified, coherent, adequately developed essay. But an objective test, the test chosen by the overwhelming majority of test users, does not test these abilities. If a test does not require the student to write, it is an incomplete test of the content of a composition course.

There are composition teachers for whom no objective test of writing ability is acceptable. Their demand that a writing sample be part of any test of student composition ability is understandably reasonable if the test is used for student exemption from or placement in courses which have as their principal activity writing essays. Composing an essay requires originality, thought, and knowledge as well as a background in "correctness" and felicities; rhetorical and stylistic choices should be made. Teachers demanding either an essay test or an objective test having an essay section claim the essay is a direct, not oblique or associational, measure of the several components of composition ability, components not tested for by objective questions. Students enter college to acquire or enhance their ability to compose, not just their ability to proofread and edit.

Those rejecting essay tests claim such tests lack reliability; that is, readers will disagree on the quality of a student's essay. Low correlations of essay test scores with objective test scores as well as with course grades are another of their reasons for rejecting essay tests.

Lack of reliability and low correlations can be overcome. Two

major works by ETS-CEEB researchers--Godshalk, Swineford, and Coffman's The Measurement of Writing Ability and Diederich's Measuring Growth in English--show how acceptable reliability coefficients and correlations can be established among readers, between essay scores and objective test scores, and between essay scores and course grades.

Test scores are correlated with other test scores and course grades to aid in the prediction of the student's classroom performance. Since objective tests usually do correlate more highly than essay tests with other objective tests and course grades, they are usually better predictors. However, the correlation is typically only a moderate 0.50 ($r = 0.50$). To help make a correlation meaningful, the coefficient of determination can be computed. The coefficient of $r = 0.50$ is 25 percent. A 0.50 correlation of test scores and course grades accounts for only 25 percent of all the variables (student ability, test validity, class attendance, test anxiety, and many others) that affect the interdependence of whatever the test measures with whatever the grade measures. Seventy-five percent of the variables are not accounted for.

Computing the index of forecasting efficiency can also help make the correlation meaningful. When $r = 0.50$, the index of forecasting efficiency is 13 percent. By knowing the student's test score, the teacher's ability to predict the student's classroom performance is 13 percent greater than by chance. Eighty-seven percent of the time any random process like flipping a coin would just as accurately predict the student's performance. Impressive as correlations are for anyone intimidated or befuddled by statistical data, they are no more or less valuable as a criterion for test acceptance than content analysis, score gains, and the like.[7]

Score gains studies can complement content analyses when content

validity is being established. If the students make statistically significant gains on equivalent forms of a test (a pre-test and post-test given before and after taking a course), the test might have content validity. It is possible, however, that the test is not measuring the effect of instruction or an increase in the student's knowledge but rather intellectual maturation. And, despite appearances, even if the students do not make statistically significant gains, the test still might have content validity. Perhaps there is no gain to be measured; perhaps teachers do not teach and students do not learn in college composition courses. Several contrary inferences can be drawn from score gains studies.

Every college English department should conduct a score gains study when validating a test. The composition ability of all students should be tested by one form of the test before they take any composition courses. After one year of college attendance, samples should be drawn from three groups: students who have taken no courses; students who have taken one course; students who have taken two courses. The three samples should each be made up of students equal in composition ability as measured by the pre-test. Theoretically, after they have been given the post-test, those who have taken no composition courses will show no gains. Those who have taken one course will have scores statistically higher than those who took no courses. Those who have taken two courses will have scores statistically higher than the scores for either of the other two samples.

Score gains will give credibility to claims for content validity; however, the qualifications made earlier about such studies must be acknowledged. Depending on the results of studies correlating scores and grades, the scores students earn after instruction should be nearly

the same scores students seeking exemption or advanced placement would earn.

The easily followed procedure outlined above is described in detail in textbooks on educational statistics. But if a department is reluctant to conduct its own study, colleagues teaching courses in statistics or educational measurements will often volunteer to conduct the study to give their students field experience. Also, conducting such a study is the province of any college's office of institutional research. In addition, there are testing services that will conduct such research for colleges using their tests.

III

Cutting Scores

All basic information about a test such as item analysis, correlations, and score gains should be included in the testing service's informational booklets. Unfortunately, however, such information is not always given. The formula for converting scale scores into raw scores is another noticeable omission. The scale scores for standardized tests are converted formula scores. The formula score is determined by subtracting a fraction of the wrong answers from the raw score. The raw score is the number of right answers. To accurately interpret the scale scores, the English department faculty must have a copy of the testing service's manual on scale and formula scores and must convert the scale scores into raw scores. Since such manuals are denied or given only reluctantly to test users by the testing services, the users must make an extraordinary effort to acquire them.

Converting scale scores back into raw scores can be embarassing for any department that has itself arbitrarily chosen or accepted without question a testing service's recommendation for cutoff scores.

If, for example, the department has made the 25th percentile and its corresponding scale score the cutoff for placement in regular sections, it may be chagrined to learn that answering only 30 percent of the test items correctly, but none incorrectly, will place a student in a regular section. Despite knowing slightly less than one-third of the test material, the student is placed in a regular, not remedial, section of freshman composition. To continue the example, if a scale score corresponding to the 75th percentile will place the student in an honors section, the student will need to answer 60 percent of the questions correctly and none incorrectly. But by raising the percentage to 69, approximately two-thirds of the test material, the student will raise his or her scale score to the impressive 90th percentile and be exempted from any formal composition instruction. As disconcerting as it may be for the department to learn just how much of the test material the student must know to earn a particular scale score/percentile rank, learning the standard error of measurement for the test may be equally unsettling.

Standardized test scores are not unequivocal. This fact is substantiated by the test's standard error of measurement. A student's "true" score on any test is not the score he or she earns after taking the test once. His or her "true" score ranges from one standard error of measurement above to one standard error of measurement below his or her reported score. Two out of three times the student will earn a score in this range. For example, if the standard error of measurement on a test is 6 and the student repeats the test, two out of three times his or her score will range from 6 points above to 6 points below his or her reported score. Within this range is his or her "true" score. One out of three times it would be above or below the 12 point range.

Scores are approximations, and they are approximations having considerable range. The table below gives an example of students' scale scores and the range of their "true" scores for a test having a mean of 58 and a standard error of 6. As shown below, the performance of a student ranking in the top third will, two out of three times, actually range from average to the top one-fifth. One out of three times it can be below average or above the top fifth. Conversely, the performance of the student scoring near the bottom third will, two out of three times, range from near the bottom quartile to average. One out of three times it could be in the bottom quartile or above average.

Scale Score	Scale Score Percentile	"True" Score Range	"True" Percentile Range
64	66th	58-70	50th-81st
58	50th	52-64	44th-66th
52	38th	46-58	28th-50th

The converging and overlapping of scores and percentiles illustrate why standardized test scale scores are not absolute, infallible indicators of a student's "true" achievement.

Implicit in a testing service's recommendation or a department's selection of a particular scale score/percentile rank for the cutoff for exemption from a course is, first, the assumption that the student who scores at the designated score/percentile has mastered the content of the course, and, second, the prediction that the student would earn a passing grade. Before accepting any score/percentile as a cutoff, the department must correlate test scores and course grades. Prediction of grades followed by exemption when based on scale scores that have not been correlated with grades is impossible. Within the correlations the scale scores/percentile ranks tell the test user only how the student

has done in relationship to all of the other students who have taken the test. With the correlations, as discussed earlier, the test user has a better, although limited, understanding of the relationship of the content of the test with the content of the course.

Testing services customarily correlate test scores and course grades in their pilot studies on the validity of their tests, and they usually report their findings to the test users. Also, institutional researchers will report on the correlations for students at their colleges and universities in professional journals. However, even if the testing service includes correlation studies in its informational booklets, and even if institutional researchers have published their studies, every English department must correlate the composition grades and test scores of the students attending its institution. Differences among types of institutions, their location, their students, and their course offerings all affect correlation studies. The correlations found at a community college in the South could differ dramatically from those for a small, private liberal arts college on the West Coast or a large, Midwestern state university. Whether the department conducts its own study, has it conducted by colleagues, the office of institutional research, or the testing service itself, the correlations will help establish test validity and appropriate cutoff scores for exemption and placement.

IV

How student writing ability should be measured for exemption, placement, or certification of competence is a major curriculum issue affecting every college student. Any decision to use a standardized test as one of the means cannot be based entirely on the testing

service's claims for its test or on research generated outside the institution. Each test has its weaknesses and limitations. But since it is unlikely that the testing service will acknowledge them all, their detection is incumbent on knowledgable, responsible English department faculty. To make best use of the objective test, the faculty must learn how its norms were established and what the test scores mean. In-house score gains studies should be conducted. Cutoff scores must not be arbitrarily chosen or ignorantly based on the testing service's recommendations. The department must correlate its students' scores and grades to help identify appropriate cutoff scores.

The faculty's refusal to use a standardized test as its sole measure will be buttressed by their knowledge of the test's raw scores and standard error of measurement. The department will decide how much of the test material the student must know before choosing a scale score to aid in exemption, placement, or certification. The department will recognize that the scale score is only an approximation and is not to be used with absolute certainty and rigidity. The department should anticipate a moderate correlation like $\underline{r} = 0.50$ between post-course test scores and course grades which will argue for using the test as a supplementary, not exclusive, means of determining student composition ability.

Through careful selection the department can choose a test that will help them make more confident decisions about their students' literacy. Almost inevitably the objective test they select will be essentially a test of the students' knowledge of standard edited English and ability to proofread and edit. The difficulty of the test and its emphasis on this knowledge and ability should be commensurate with that of the department's courses. But even if commensurate, the objective test

should be complemented by an essay which will reveal student abilities and knowledge excluded from the objective test. An essay is the singular example of a student's writing ability, despite its low correlations with course grades and its low reliability coefficients. Unlike the objective test which is shards, meaningful but still bits and pieces, the essay is an individually created, whole artifact.

As for those low correlations and coefficients, the department wanting to raise them can in its own workshop establish standards and identify samples of writing they would judge to be the essays of students deserving exemption or advanced placement, or needing remediation. They can abstract from their own students' essays the criteria they use to assign grades in their own courses at their own institution. Combining the student's performance on the objective test with the quality of his or her essay will provide the faculty with an excellent means of identifyingthe student's needs.

Notes

[1]Urban Whitaker, "Credit by Examination at San Francisco State College," The College Board Review, No. 83 (Spring 1972), p. 12, 16.

[2]Jerilee Grandy and Walter M. Shea, The CLEP General Examinations in American Colleges and Universities, Princeton, New Jersey: ETS-CEEB, 1976, p. 10.

[3]Edward Caldwell, "Analysis of an Innovation (CLEP)," Journal of Higher Education, 44 (December 1973), 701.

[4]Robert S. Rudolph and Richard M. Summers, "The CLEP General Examination in English Composition: What Does It Really Test?" Freshman English News, 6 (Spring 1977), 3-4.

[5]CLEP Scores: Interpretation and Use (Princeton, New Jersey: ETS-CEEB, 1976), p. 10.

[6]J. Andrew Archer and Harry C. Nickens, "Credit by CLEP: A Disconcerting Look at a Good Idea," Community/Junior College Research Quarterly, 1 (January-March 1977), 181.

[7]Correlations, their meaning, and value are explained in my article "Standardized Tests of Writing Ability: A Primer," Freshman English News, 10 (Fall 1981), pp. 20-21.